基于开放域抽取的多文档概念图构建研究 *

盛泳潘 [1]，付雪峰 [2]，吴天星 [3]

(1. 电子科技大学 计算机科学与工程学院，成都 611731; 2. 南昌工程学院 信息工程学院，南昌 330099; 3. 东南大学 计算机科学与工程学院，南京 211189)

摘　要：在信息过载的背景下，如何从拥有共同主题的多篇文档中挖掘并组织核心概念及其语义连接已成为当前开放式信息抽取任务中的一项重要挑战。为此，提出了一个基于开放域抽取的多文档概念图构建模型。首先基于预定主题挖掘主题词，通过改进的 TF-IDF 算法对文档进行排序；然后通过共指消解、篇章权重计算、开放域抽取等一系列的方法从多篇文章中抽取出大量具有事实表达能力的三元组实例。为去除开放域方法本身的噪声以及提升信息抽取的准确率，提出一种事实过滤算法。通过该算法可有效提取置信度高且具有良好语义兼容性的显著事实知识集合，并构成多个概念子图。最后，将不同子图中等价的概念以及关系进行合并，形成一张具有主题表达能力的连通概念图。通过在 Signal Media 新闻数据集上进行验证，实验结果表明，所提出的模型能够跨文档挖掘并有效组织与特定主题相关的关键信息，形成的概念图在主题概念覆盖率、事实知识的兼容性等指标上均取得了较好的效果。除此之外，该模型对于自动文档摘要的应用也具有重要的参考价值。

关键词：开放域抽取；多文档；概念图构建

中图分类号：TP391　　doi: 10.19734/j.issn.1001-3695.2018.05.0454

Multi-document conceptual graph construction research based on open domain extraction

Sheng Yongpan[1], Fu Xuefeng[2], Wu Tianxing[3]

(1. *School of Computer Science & Engineering*, *University of Electronic Science & Technology of China*, *Chengdu* 611731, *China*; 2. *School of Information Engineering*, *Nanchang Institute of Technology*, *Nanchang* 330099, *China*; 3. *School of Computer Science & Engineering*, *Southeast University Nanjing* 211189, *China*)

Abstract: In the background of information overload, this is challenging to mine and organize meaningful concepts and their semantic connections from a set of related documents under the same topic in open information extraction. Thus, this paper proposed a multi-document conceptual graph model based on open-domain information extraction. Firstly, documents were ranked according to the improved TF-IDF weight of extracted topic words under the predefined topics, then the model relayed on a serious of methods, including coreference resolution, weight computation, open-domain information extraction method to extract numerous representative subject-predicate-object triples from multiple documents. For filtering out the noise of open-domain information approach itself and improving the accuracy of information extraction, this paper presented a fact filtering algorithm to retain only the most salient, compatible facts as well as a form of multiple conceptual subgraphs. Finally, in combined with the equivalent concepts and relationships across different subgraphs to connect into a fully connected conceptual graph with expressive topic ability. Experiments on Signal Media dataset illustrated that the proposed model has the ability to discern and effectively group the key information corresponds to specific topics within and across documents, and formed conceptual graph outperforms state-of-the-art the algorithms in terms of the coverage rate of topic concepts as well as the compatible facts. Besides, this model also has the important significance for the automatic abstract on.

Key words: open-domain extraction; multiple documents; conceptual graph construction

0　引言

随着大数据时代的不断演进与发展，通过报纸、广播电视、互联网、微博、微信等媒体渠道发布的以及用户所创造的信息急剧增长，自由文本数据作为其中的典型代表，为揭示信息实质与语义关系起到了重要的推动作用。然而，如何从大规模、主题信息零散分布的文本集合中获取重要的主题概念以及语义关联，已成为当前信息抽取任务中的一项重要挑战。

文献[1]提出多文档摘要技术，其旨在于将多篇拥有共同主题的文章的大意提取出来，形成简练可读、易于用户理解的短

收稿日期：2018-05-23；**修回日期：**2018-08-03　　**基金项目：**国家自然科学基金资助项目（61762063）；江西省自然科学基金资助项目（20171BAB202024）；江西省教育厅科研项目（GJJ170991）；国家建设高水平大学公派研究生项目（201706070049）

作者简介：盛泳潘，男，博士研究生，主要研究方向为知识图谱、自然语言处理；付雪峰，男（通信作者），讲师，博士，主要研究方向为符号逻辑对不精确知识的表示与推理（fxf@nit.edu.cn）；吴天星，男，博士，主要研究方向为知识图谱、语义网、数据挖掘.

文本。文献[2]基于 LexRank 算法提出多文档摘要自动化抽取方法，该方法可作为概括式摘要的典型代表。随后，文献[3]利用神经网络模型对文本中心句进行建模，实现了对摘要语句的进一步压缩。文献[4]提出了潜在狄利克雷分布（latent Dirichlet allocation, LDA）的主题模型，能够从大规模文本数据集中挖掘隐含的主题信息，但需要手动指定主题的数目。文献[5]提出了层次潜在狄利克雷分布（hierarchical latent Dirichlet allocation，HLDA）的主题模型，解决了上述缺陷。文献[6]充分利用词频、主题语句位置、主题词等特征，设计出一种自动文本摘要抽取系统。文献[7]提出了一种评论式摘要，在保留评论人总体观点的情况下，同时反映出了评论的多样性。然而自摘要的方法在主题概念覆盖率（如融合零散的主题细节信息）、准确度等方面均表现一般，优选主题概念、关联信息，融合生成流畅且有意义的多文档摘要一直以来都是信息抽取领域所研究的关键问题。文献[8]提出了开放域抽取方法，该类方法以语句依存分析为基础，能够适应无标注的非限定领域的大规模文本的开放式信息抽取任务。华盛顿大学在信息抽取领域，研究出了 KnowItAll、TextRunner、WOE、ReVerb、R2A2 等一系列具有里程碑意义的二元 OIE（open information extraction）系统[9]，其最主要的优势在于能够在兼顾上下文全局信息的同时，对语料中的二元浅层实体关系进行高效抽取。然而，由于开放域方法主要依赖于开放模板、语句的依存关系等特征，并不能较好地识别出所抽取的事实是否能够准确地表达出语句的含义，并且难以跨语句、跨文档有效的连接这些事实知识。文献[10]采用一种逻辑约束的方法，实现了跨文档组织事实知识的目的，但由于该规则仅限于有限关系的应用场景，未能保证所连接的事实表达是有意义的，并且能够覆盖重要的主题信息。

在开放式信息抽取任务中还有一类典型的方法，即开放式实体关系抽取方法。该方法主要基于以下假设：若已知实体间存在指定的语义关系，则所有包含这两个实体的句子都隐式地表达了这种关系。开放式实体关系抽取方法[11]主要是通过借助外部领域无关的知识库（如 DBPedia、Freebase、YAGO、wikipedia 文本库等），将高质量的实体关系映射到知识库语料中，然后根据文本对齐方法从中获取关系抽取训练数据（该过程也可被视为数据标注过程），并训练模型实现关系抽取任务。然而该类方法主要存在以下两个方面的问题：a）训练语料存在较多噪声；b）标注的实体关系类型有限。针对前者问题，远程监督（distant supervision）抽取方法自提出以来就受到了业内专家的普遍关注，并且取得了良好的性能。文献[12]针对传统统计模型在特征抽取过程中出现的错误、错误传播，以及深度学习方法中依靠单一词向量来学习特征的不足，提出了一套基于卷积神经网络与关键词策略相结合的实体关系抽取方法，实验表明该方法有利于提升抽取结果的准确率。文献[13]针对数据标注错误的问题，采用多示例学习的方式从训练集中抽取置信度高的训练样例来训练模型，对于算法性能的提升起到了一定的成效。针对后者问题，目前更多的实体关系抽取方法[14]尝试面向大规模的开放语料，其所包含的关系类型将更加全面。

针对开放式信息抽取任务中难以跨语句、跨文档组织事实知识信息以及标注实体关系类型较为有限这两方面问题，本文提出了一套基于开放域抽取的多文档概念图构建模型，该模型依赖于一系列 NLP（natural language processing）方法以及工具，通过概念图的形式表现出特定主题下显著的实体、概念，以及它们之间的关系，实现了跨文档挖掘并组织主题关键信息的目的，对于进一步研究该主题的发展脉络以及自动文档摘要的应用具有重要的参考价值。

1　多文档概念图构建模型

构建基于多文档语义链接的概念图模型主要包括四个主要任务，分别为文档排序、概念及关系抽取、事实过滤、合并等价概念及关系，以下将详细进行阐述说明。

1.1　文档排序

基于预定义的主题，通过 Stanford CoreNLP 系统[15]挖掘文档中的命名实体、动名词、名词、事件名称等作为候选关键词，并通过改进的 TF-IDF 算法计算它们对主题的重要程度。与传统的 TF-IDF 算法相比，该算法不仅降低了生僻词被误识为主题词的概率，而且考虑了关键词在不同主题间的分布情况。其计算公式为

$$\text{TF-IDF} = tf(w,|\mathrm{D}|) * idf(w,k,k^*) \tag{1}$$

其中：$tf(w,|\mathrm{D}|)$ 表示词频（term frequency），用于衡量关键词在特定主题所有文档中的重要程度；$idf(w,k,k^*)$ 表示逆文档频率（inverse document frequency），用于衡量关键词在所有文档中的通用程度，并不限于特定的主题。

$$tf(w,|\mathrm{D}|) = 1 + \log(\frac{c(w)}{n(|\mathrm{D}|)}) \tag{2}$$

$$idf(w,k,k^*) = -\log(1 - \frac{f(w,k)}{f(w,k) + f(w,k^*)}) \tag{3}$$

其中：w 为候选关键词；k 为特定的主题；$|\mathrm{D}|$ 表示 k 下的文档总数；$n(|\mathrm{D}|)$ 表示 $|\mathrm{D}|$ 中的单词总数；$c(w)$ 表示 w 在 $|\mathrm{D}|$ 中出现的次数；$f(w,k)$ 表示 w 在 $|\mathrm{D}|$ 中的频率；k^* 表示除 k 外的其余主题；$f(w,k^*)$ 表示 w 在 $|\mathrm{D}^*|$ 中的频率。

文档权重的计算公式如下：

$$weight(k,d) = \sum_{i=0}^{n(d_{key})} w_i^{\text{TF-IDF}} \tag{4}$$

其中：d 为 k 下的文档；$n(d_{key})$ 表示 d 中包含的关键词总数；$w_i^{\text{TF-IDF}}$ 表示 d 中第 i 个关键词的 TF-IDF 值，可根据式（1）进行计算。

1.2　概念及关系抽取

概念及关系抽取的主要任务是从同一主题下的多篇文档中抽取出大量具有事实表达能力的三元组实例，主要包括共指消解、篇章权重计算、开放域抽取三个子任务。

1)共指消解　同一篇文章中的指代类型主要表现为人称指代、指示性指代、名词短语指代以及事件指代。本文采用斯坦

福大学研发的自然语言处理的工具包——Stanford CoreNLP 系统对单篇文档中的共指代词进行替换，目的在于提高文档语句的可读性，以利于后续的开放域抽取任务。

2）篇章权重计算　TextRank 算法[16]的基本思想来源于 Google 著名的 PageRank 算法，通过将文本切分为若干语义单元并建立图模型，利用投票机制对文本中的重要成分进行排序，可用于单篇文档的关键词提取、自动摘要等任务。 本文采用 TextRank 算法计算文档中不同语句的得分，并将高分语句作为文档的主题句。

3)三元组实例抽取　传统的信息抽取模式需要限定领域以及语义单元的类型，无法应用于未预先定义概念关系类型的自由文本语料。因此，可通过华盛顿大学研发的新一代 OLLIE（open language learning for information extraction）系统[9]对文档主题句中的二元关系进行抽取。抽取的三元组实例可表示为（subject, predicate, object）的形式。其中，subject、object 表示不含嵌套结构的两个实体或概念；predicate 表示它们之间的关系，主要以不含嵌套关系或修饰短语的动词及动词短语为主，对于复杂的长句，通过 OLLIE 系统会抽取出一个或多个具有不同置信度的关系对。下面通过三个例句进行解释说明。

例句 1：82 percent of leaned Democrats say Registered voters'd support a clear Clinton, while 76 percent of leaned Republicans say Registered voters'd back a clear Clinton vs. Trump, were Registered voters the party nominees.

F1:　0.97 (76 percent of leaned Republicans; say; Registered voters'd back a clear Clinton vs. Trump)

F2:　0.94 (82 percent of leaned Democrats; say; Registered voters'd support a clear Clinton)

在例句 1 中，通过 OLLIE 系统可抽取出置信度为 0.97 的三元组实例 F1 和置信度为 0.94 的三元组实例 F2。置信度越高，说明三元组实例所表达的事实知识越准确。

例句 2：That compares to a clear Clinton lead among all adults, 51-39 percent, indicating her broad support in groups that are less apt to be registered to vote, such as young adults and racial and ethnic minorities.

F3:　0.45 (That; compares; to a clear Clinton lead among all adults)

F4:　0.90 (51-39 percent; indicating; her broad support in groups)

F5:　0.67 (groups; are; less apt to be registered to vote)

例句 3：The hypothetical contest compares to a clear Clinton lead among all adults, 51-39 percent, indicating Hillary Clinton broad support in groups that are less apt to be registered to vote, such as young adults and racial and ethnic minorities.

F6:　0.95 (The hypothetical contest; compares; to a clear Clinton lead among all adults)

F7:　0.90 (51-39 percent; indicating; Hillary Clinton broad support in groups)

F8:　0.92 (groups; are; less apt to be registered to vote)

例句 3 是对例句 2 进行共指消解处理后的结果，原句中划线部分的 That 被替换成了 The hypothetical contest，her 被替换成了 Hillary Clinton。相应的，That 指代所对应的三元组实例由 F3 变成了 F6，其置信度由原来的 0.45 提升到了 0.95；her 指代所对应的三元组实例 F4，经共指消解处理后的置信度由 0.67 提升到了 0.92。

1.3　事实过滤

OLLIE 系统易受依存分析错误的影响，产生无信息量或错误的三元组实例。与此同时，多篇文档中重复语义的语句会产生一定比例的、冗余的三元组实例。针对以上两方面的问题，本文提出一套事实过滤算法，目的是为了过滤掉与主题核心内容无关并且低置信度的候选三元组实例，只保留那些置信度较高且具有良好语义兼容性的显著事实知识信息。该算法将三元组实例的过滤问题转换为整数规划问题，目标方程及相应的约束条件如下所示：

$$\max_{x,y} \alpha^T x + \beta^T y \tag{5}$$

$$s.t.\quad 1^T y \le n_{\max} \tag{6}$$

$$x_k \le \min\{y_i, y_j\} \tag{7}$$

$$\forall i < j, i, j \in \{1, \ldots M\} \tag{8}$$

$$k = (2M - i)(i-1)/2 + j - i \tag{9}$$

$$x_k, y_i \in \{0,1\} \forall i \in \{1,\ldots,M\}, k \tag{10}$$

其中：$x \in \mathbb{R}^N$；$y \in \mathbb{R}^M$；$N = (M+1)(M-2)/2+1$；$T = \{t_1,\ldots,t_M\}$ 为包含 M 个元素的三元组实例集合；$t_i, t_j \in T$（$i, j \le M, i \ne j$），表示集合中的任意两个三元组实例；y_i 为 t_i 的指示变量，即：如果 y_i 为真，t_i 被保留；x_k 同样为指示变量，表示 t_i 与 t_j 之间的兼容性，即：如果 x_k 为真，这时 y_i=1，y_j=1，t_i 和 t_j 均被保留；β_i 表示 t_i 所表述事实的置信度；$n_{\max}$ 为概念图中的三元组实例个数，该值可由用户进行设置，在算法所生成的三元组实例集合中会包含不大于 $n_{\max}$ 个数的三元组实例。

α_k 表示 t_i 和 t_j 的语义兼容性,其计算公式如下:

$$\alpha_k = sim(t_i, t_j) = \gamma \dot{s}_k + \eta \dot{l}_k \tag{11}$$

其中：$\dot{s}_k$ 表示 t_i 和 t_j 的语义相关性，主要通过 ADW（align disambiguate and walk）模型[17]进行计算；$\dot{l}_k$ 表示 t_i 和 t_j 的字面相似度，主要通过 Levenstein 距离公式进行计算；γ 为比例系数，表示 $\dot{s}_k$ 在 α_k 计算中所占的比例；η 表示 $\dot{l}_k$ 所占的比例，并且 $\gamma + \eta = 1$。

为了减小计算负载，方法中引入了滑动窗口机制，即随着滑动窗口的移动，每次只比较窗口内未重复计算的三元组实例，计算复杂度由 $O(M^2)$ 降为 $O(\Delta WM)$。其中：$\Delta W = 2W - step - 1$；

w 为窗口大小；$step$ 为滑动步长。

1.4 合并等价概念及关系

当前任务的难点在于概念指称的多样性，以及对概念关系的描述可能存在较大噪声。因此，本文提出以下规则来合并多个概念子图中等价的概念以及关系。

规则 1 同义概念具有等价性。同义概念在词汇结构上具有明显的特征，例如 Billionaire Donald Trump、Donald Trump、Donald John Trump、Trump 都指向同一人物。对于命名实体，可借助于搜索引擎强大的实体链接能力，检查它们是否能够准确链接到同一指称对象。

规则 2 相似的概念具有等价性。主要采用 ADW 模型[14]，它依赖于 WordNet 词典，通过执行随机游走可获得概念对应的语义指纹，并通过 Cosine、Weighted Overlap、Top-k Jaccard 三种方法计算两个概念指纹间的相似性。

规则 3 语义重合的概念具有等价性。主要采用文献[18]提出的语义重合度计算公式。该度量方法依赖于 WordNet 的分类结构，通过将两个概念到根节点的路径长度转换为信息量进行计算。

通过 OLLIE 系统抽取的关系描述中存在长尾关键词以及噪声。为保证识别的准确度，等价关系的标注工作主要由 NLP 专家标注者来完成，具体包括：a）标注者根据背景知识、关键词重叠度、连接概念的一致性等依据，在多个概念子图中标注若干等价关系对，并结合它们所连接的概念（等价关系中至少有一个概念是相同或等价的）完成合并任务；b）根据文献[1]中所提出的：一个良好的概念图最多不应超过 25 个概念，并且应具有连通性，因此，如果最大概念子图所连接的概念数量未达到 25 个，且多个子图未形成连通，允许标注者依据背景知识定义新的语义关系标签（最多不超过 3 个），使子图间的概念形成连接，构成一张连通的概念图。新的关系标签可依据概念间的基本关系进行定义，如施事关系、拥有关系、目的关系、主观关系等。为难免单个标注者在合并等价概念以及合成语义关系的过程中所产生的片面性认知，上述任务至少需要由两个以上的专家标注者配合完成，其标注结果将通过 Kappa 系数进行一致性检验。

2 实验验证与分析

2.1 实验数据

Signal Media 收集的新闻报道记录了 2015 年 9—10 月期间通过 Reuters 发布的国际热点新闻（http://research.signalmedia.co/newsir16/signal-dataset.htm），共计 1 000 000 篇英文文档，其中包含 734 488 篇新闻、265 512 篇博客，每篇文档平均拥有 39 个句子、1 266 个单词。在 DUC（document understanding conference）标准语料[19]中，同一个主题下大约包含 25~40 篇文档，而本实验随机抽取 10 000 篇作为研究语料，其中包含 734 篇新闻（占 73.4%）和 266 篇博客（占 26.6%），语料共分为 Syria refugee crisis（叙利亚战争危机）、Iran nuclear（伊朗核问题）、Volkswagen scandal（大众汽车丑闻）、United states presidential election（美国总统选举）、Sino-Soviet cooperation（中苏合作）五个主题，根据本文 1.1 节中的式（4），在每个主题中选取前 100 篇文档，在其中随机选取文档进行分析，生成如下规模的数据集：5，15，25，35，45，55，65 和 75（单位：篇），并将数据集命名为：D_1^i，D_2^i，D_3^i，D_4^i，D_5^i，D_6^i，D_7^i，D_8^i（i 为指定主题，$1 \le i \le 5, i \in N^+$），通过分析上述数据集来测试本文模型的性能。实验数据集的具体情况如表 1 所示。

表 1 实验数据集

Table 1 Datasets used in experiments

主题名称	文档个数	单文档大小	标准差	候选主题词个数
叙利亚战争危机	100	1715±614	1.51	654
伊朗核问题	100	1069±537	0.56	429
大众汽车丑闻	100	999±326	0.54	598
美国总统选举	100	1175±207	1.26	772
中苏合作	100	768±122	0.32	280

2.2 评价指标

下面将从主题概念覆盖率、概念图连通性、概念图可读性、模型运行时间以及对比算法五个方面对本文所提出的概念图模型进行全面分析，其中涉及的评价指标如下：

a）主题概念覆盖率。表示正确抽取的主题概念在概念图中所占的百分比，计算公式为

$$C_{theme} = \frac{n_{theme}}{n_{concept}} \tag{12}$$

其中：$n_{concept}$ 为概念图中的概念总数；n_{theme} 为通过随机森林算法计算得到的主题概念的数量。

b）Kappa 系数。用于对标注结果的一致性检验，计算公式为

$$K = \frac{P_0 - P_e}{1 - P_e} \tag{13}$$

其中：P_0、P_e 分别为不同标注结果的观察一致率与机遇一致率；$P_0 - P_e$ 为实际一致率；$1 - P_e$ 为非机遇一致率。

c）ROUGE 评测标准。一种基于召回率的相似性度量方法，主要包括 ROUGE-N、ROUGE-L、ROUGE-S 等评价指标。

ROUGE-N 表示基于 N-gram 的共现性统计，其准确率 $ROUGE-N_P$、召回率 $ROUGE-N_R$、F 值 $ROUGE-N_F$ 的计算公式分别为

$$ROUGE-N_P = \frac{\sum_{S \in \{ct_summary\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ct_summary\}} \sum_{gram_n \in S} Count(gram_n)} \tag{14}$$

$$ROUGE-N_R = \frac{\sum_{S \in \{gt_summary\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{gt_summary\}} \sum_{gram_n \in S} Count(gram_n)} \tag{15}$$

$$ROUGE-N_F = \frac{2 \times ROUGE-N_P \times ROUGE-N_R}{ROUGE-N_P \times ROUGE-N_R} \tag{16}$$

其中：n 表示 N-gram 词元共现的长度；$gram_n \in ct_summary$ 表示在生成摘要中出现的 N-gram，$gram_n \in gt_summary$ 表示在标准摘要中出现的 N-gram；$Count_{match}(gram_n)$ 表示在生成摘要和标准摘要共现的 N-gram 数目。

ROUGE-L 表示基于最长公共子序列（LCS）的共现率统计，其准确率 P_{lcs}、召回率 R_{lcs}、F 值 F_{lcs} 的计算公式分别为

$$P_{lcs} = \frac{\sum_{i=1}^{r} LCS(c_i, C)}{n} \tag{17}$$

$$R_{lcs} = \frac{\sum_{i=1}^{r} LCS(c_i, C)}{m} \tag{18}$$

$$F_{lcs} = \frac{(1+\beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \tag{19}$$

其中：$LCS(c_i, C)$ 表示参考摘要与系统摘要中 LCS 的并集；$\beta = \frac{P_{lcs}}{R_{lcs}}$ 为衡量 P_{lcs} 与 R_{lcs} 之间重要度的平衡系数；m 和 n 分别为系统摘要和参考摘要包含的语句数目。

ROUGE-S 表示基于长度顺序子序列的共现率统计，其准确率 P_{pair}、召回率 R_{pair}、F 值 F_{pair} 的计算公式分别为

$$P_{pair} = \frac{pair(x, y)}{Comp(n, 2)} \tag{20}$$

$$R_{pair} = \frac{pair(x, y)}{Comp(m, 2)} \tag{21}$$

$$F_{pair} = \frac{(1+\beta^2) R_{pair} P_{pair}}{R_{pair} + \beta^2 P_{pair}} \tag{22}$$

其中：$pair(x, y)$ 表示词对 (x, y) 共现匹配的数量；$\beta = \frac{P_{pair}}{R_{pair}}$ 为衡量 P_{pair} 与 R_{pair} 之间重要度的平衡系数；$Comp(m, 2)$ 表示系统摘要中词对的组合数；$Comp(n, 2)$ 表示参考摘要中词对的组合数。

2.3 主题概念覆盖率分析

2.3.1 事实过滤算法中 W、$step$、γ、η 的取值

从理论上来说，滑动窗口值 W、滑动步长 $step$ 的取值越大，可利用的三元组实例的上下文信息就越多。但本文所提出的模型主要关注于滑动窗口内的三元组实例的语义兼容性特征。因此，如果上述参数设置过大，反而会使三元组实例集合整体的语义兼容性降低，同时也会影响模型的运行效率、造成资源的浪费；如果参数的值设置过小，很有可能获取不到足够多的有用信息。

γ 值决定了语义相关性因素在衡量两个三元组实例的语义兼容中时所占的比例，η 值决定了字面相似度因素在衡量两个三元组实例的语义兼容性中所占的比例，并且 $\gamma + \eta = 1$。

在事实过滤算法中，对所有结果进行统计分析，发现表 2 中的参数取值可使抽取的三元组实例集合的语义兼容性达到最高，其中 n_{max} 表示由用户指定的概念图中的三元组实例的个数；$step$ 的取值依赖于滑动窗口值 W。

表 2 事实过滤算法中参数的最佳取值

Table 2 Best parameters in fact filtering algorithm

滑动窗口值	滑动步长	比例系数	比例系数
W=1/4 n_{max}	$step$=1/2 W	γ=0.8	η=0.2

2.3.2 主题概念覆盖率

根据本文 1.1 节的式（1），计算实验所设置的五个主题中的不同候选主题词对于主题的影响程度，在每个主题下选择前 200 个具有高 TF-IDF 值的概念作为主题概念，并以表 3 中所述特征，通过随机森林算法训练得到一个二分类器，通过二分类器计算由本文模型所构建的概念图中每个概念的 Gini 系数。当概念的 Gini 系数大于 δ 时，则判定其为主题概念；否则判定其为非主题概念。通过训练使模型的准确率达到 92.3%。此时算法参数设置如下：阈值 δ=0.5，概念图中的三元组实例的总数 n_{max}=20，子树的数量 n_estimators=50，划分时的最大特征数 max_features=6，决策树的最大深度 max_depth=3，内部节点划分所需最小样本数 min_samples_split=4，叶子节点最少样本数 min_samples_leaf=2。

表 3 随机森林算法的分类特征

Table 3 Features used for random forests for classification

特征	含义	取值范围
Frequency	词频	[0, 1]
Is_MatchLength	当前概念是否少于 5 个或多于 12 个字符	0 or 1
Is_InAbstract	是否在自动抽取式摘要[17]中出现	0 or 1
Is_MatchLanguagePattern	是否满足 premodifier+ headword+ postmodifier 的英文语言模式	0 or 1
Is_NER	是否为命名实体[11]	0 or 1
Synonyms	当前概念的同义词占概念图中所有概念的比例	[0, 1]

在实验设置的五个主题下，根据不同的文档数据规模 N（D_1^i，D_2^i，D_3^i，D_4^i，D_5^i，D_6^i，D_7^i，D_8^i，$1 \le i \le 5, i \in N^+$），主题概念覆盖率 C_{theme} 的变化情况如图 1 所示。

从图 1 中可以看出，在不同主题下，随着文档数据规模的增加，主题概念的覆盖率均呈现下降的趋势，其主要原因是因为文档数量的增加使得主题信息分布的离散程度变大，OLLIE 系统的抽取精度也有所下降，在 DUC 标准语料规模 D_3 - D_4 下，综合来看，本文模型能够保留住 84%的主题信息，说明模型的精度和泛化能力较好。对于不同的主题，由于其下所包含的文档的大小、候选主题词的粒度均有所不同，例如在“美国总统选举”主题下，其所包含的单文档大小、候选主题词个数均为最多，在 D_1 规模下，C_{theme} 达到了 92%；在最大数据集规模 D_8 下，C_{theme} 为 80%。相比之下，“中苏合作”主题下的单文档大小与候选主题词个数均为最少，在 D_1 规模下，C_{theme} 达到了 84%；在 D_8 规模下，C_{theme} 下降到了 68%。“伊朗核问题”主题与“大众汽车丑闻”主题下所包含的文档信息最为类似，C_{theme} 随文档规模 N 的变化情况也极为类似，由此说明本文模型与上述两个

因素具有较强的相关性，在文档数据规模为 DUC 标准语料规模，候选主题词数量足够多的情况下，基于当前的分类器，模型能够发挥最好的性能。

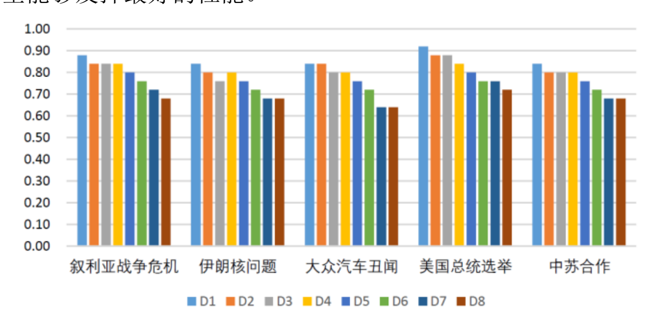


图 1　主题概念覆盖率在不同文档数据规模下的变化情况

Fig.1　Variations of a coverage rate of topic concepts in documents with different scales

2.4　概念图连通性分析

针对实验设置的五个主题，发现通过事实过滤算法而得到的概念子图集合中平均只有 47%的三元组实例易于连接，即它们的头概念或尾概念中至少有一个具有相同的形式。因此，对等价概念以及关系的判定、标注工作将直接影响最终所生成的概念图的连通性。本实验以 DUC 标准语料库规模为参照，根据本文模型选择 D_3^i（$1 \le i \le 5, i \in N^+$）规模的数据生成概念图，其中事实过滤算法中的参数设置如下：n_{max}=20，γ、η、*step*、w 选取最佳取值，分析的结果如表 4 所示。与此同时，采用 Robert Tarjan 提出的 Tarjan 算法检查所生成的概念图的强连通性，分析结果如表 5 所示。

表 4　概念子图分析结果

Table 4　Analysis results of conceptual subgraphs

主题名称	概念子图数	等价概念对	新的语义关系标签	Kappa 值
叙利亚战争危机	3	10	3	0.81
伊朗核问题	4	3	2	0.86
大众汽车丑闻	3	5	3	0.83
美国总统选举	3	11	3	0.88
中苏合作	5	4	3	0.84

从分析结果可以看出，每个主题下平均拥有三个以上的概念子图需要进行合并处理，并且这些子图大多为强连通图，同时专家标注者在进行关系合成时表现出了较好的一致性，说明本文模型在保证概念图的总体连通性方面表现较好。最大强连通分量置信度比指的是最大强连通分量中三元组实例的置信度（置信度以 OLLIE 系统的抽取结果为依据）的总和占合成的概念图中所有三元组实例置信度的比例。该比例越高，说明强连通分量的事实表达能力越强，合成的概念图的总体连通效果越好。例如，“美国总统选举”主题与“大众汽车丑闻”主题下虽然具有相同的强连通分量个数，但是由于后者中的最大强连通分量置信度比较低，只有 10%（即包含相互连通且具有较强事实表达能力的三元组实例的数量较少），说明其总体连通效果不如前者。最大强连通分量中的主题概念覆盖率表示主题概念在概念图连通分量中的分布情况，与其他主题相比，在“美国总统选举”主题下，最大强连通分量中的主题概念覆盖率达到了 84%，基本接近于概念图在当前数据规模下的主题概念覆盖率，说明本文模型在该主题下的连通效果达到最佳状态。

表 5　概念图连通性分析结果

Table 5　Analysis results of connectedness of conceptual graphs

主题名称	强连通分量	最大强连通分量置信度比	主题概念覆盖率
叙利亚战争危机	2	82.3%	0.72
伊朗核问题	4	32.6%	0.68
大众汽车丑闻	2	10%	0.80
美国总统选举	2	92%	0.84
中苏合作	4	36.2%	0.64

2.5　概念图可读性分析

一个有效的概念图除了需要涵盖足够多的主题信息外，还应具备良好的可读性。为了验证根据本文模型所得到的概念图中主题概念间的语义兼容度以及概念图整体的信息可读性。本实验以文本摘要的评测指标 ROUGH 作为评估标准，根据 2.3.2 小节的分析结果，从主题覆盖率最高的“美国总统选举”主题下选择 D_3^4 规模的文档数据生成概念图，其中事实过滤算法中的参数设置如下：n_{max}=20，γ、η、*step*、w 选取最佳取值。

为使生成的概念图满足摘要的形式，两个概念标注者需将概念图中的事实信息的顺序依次进行调整（Kappa=0.89），最终形成摘要。对于上述规模的文档数据集，依靠领域专家进行分析，总结生成抽取式摘要的方法显然是不现实的。因此，首先将实验中的所有文档做共指消解处理，以进一步提高语料的可理解性；然后，通过文献[20]中提出的经典抽取式摘要算法生成标准摘要；最后，将生成摘要与标准摘要进行比较，评测结果如表 6 所示。

表 6　ROUGE 标准评测比较

Table 6　Comparison results in ROUGE criteria

评测标准	Avg_Precision	Avg_Recall	Avg_F1
ROUGE-2	0.643	0.438	0.521
ROUGE-L	0.517	0.259	0.346
ROUGE-S	0.344	0.384	0.362

从比较结果来看，ROUGE-2 标准的准确率最高，说明通过本文模型提取得到的事实信息具有良好的文本覆盖率，并且能体现出一定的顺序特征，满足可读性的基本要求。但是对于 ROUGE-L 与 ROUGE-S 标准，本文模型表现较为一般，其主要原因是因为受限于上述文档数据规模以及 OLLIE 抽取系统本身的噪声，模型很难反映出句子级别的事实顺序特征。

2.6　模型运行时间分析

本实验以 DUC 标准语料库规模为参照，根据 2.3.2 小节的分析结果，从主题覆盖率最高的“美国总统选举”主题下选择不同规模的文档数据集 N（$D_j^4 \in N$，$1 \le j \le 8, j \in N^+$）生成概念图，其中事实过滤算法中的参数设置如下：n_{max}=20，γ、η、*step*、w 选取最佳取值。将模型中的文档排序、概念及关系抽

取、事实过滤、合并等价概念及关系四个主要任务简记为任务 1、2、3、4，分析不同任务的平均运行时间在当前文档数据集上的变化情况，如表 7 所示。

表 7　本文模型在不同任务上的平均运行时间比较　　/s

Table 7　Comparison of mean running time ofour model for different tasks　　/s

文档数据集 N	任务 1	任务 2	任务 3	任务 4
D_1^4	43.3	61.5	90.4	79.7
D_2^4	59.2	91.6	131.2	87.7
D_3^4	79.5	139.6	181.6	87.9
D_4^4	98.6	170.7	228.9	87.4
D_5^4	113.5	203.1	267.6	86.6
D_6^4	137.4	228.4	303.3	87.7
D_7^4	154.6	250.7	349.6	89.6
D_8^4	171.9	273.6	387.7	87.1

从表 7 中可以明显看出，基于上述主题，本文模型在任务 3 上的计算消耗是最多的，平均占据了模型运行时间的 60%，究其原因主要是因为在该任务中需要通过 ADW 模型[17]计算当前滑动窗口内的三元组实例的语义兼容性，ADW 模型依赖于 WordNet 庞大的词典来获得概念所对应的语义指纹信息，导致计算效率显著降低；任务 1 的运行时间主要取决于文档数据的质量，总体来说，由于本实验选取的文档数据集具有较高的主题覆盖率，所以在该任务上的运行时间并不会因 N 的增加而产生跳跃式变动；任务 2 主要取决于 OLLIE 系统的性能，其运行时间随 N 的增加呈现稳定的增长；任务 4 的运行时间基本保持稳定。

随着文档数据规模的增长，本文模型在各个任务上的运行时间随任务的计算复杂度 $O(n)$ 呈线性增长，当文档数据规模达到 D_8^4 时，其运行时间并没有出现跳跃式增长，依然保持在用户可接受的范围之内。综上所述，本文模型运行稳定，在适应数据增长方面具有良好的性能。

2.7　对比算法分析

本实验以 DUC 标准语料库规模为参照，选择 D_3 规模的文档数据（其中包含五个主题文档集合，即 D_3^i，$1 \le i \le 5$，$i \in N^+$）进行测试分析，将本文模型与代表性方法进行对比，给出它们在主题概念覆盖率、等价概念对数量、新的语义关系标签数量、强连通分量个数、事实知识的兼容性等六个测试指标上的平均值，对比结果如表 8 所示。其中，抽取事实知识的语义兼容性可通过式（11）进行计算。文献[20]可视为抽取式摘要中的典型方法，其主要通过单调亚模函数建立目标函数，将多文档中主干语句的选择转换为优化问题，然后利用贪婪算法求得最优解，并且取得了较好的性能。本实验将通过该方法生成标准摘要。

1）文献[21]　Stanford OpenIE 模型是 OIE（open information extraction，开放式信息抽取）中的代表方法。

2）文献[22]　主要采用主谓宾句法关系来抽取三元组实例。

3）文献[23]　基于分句依存关系局部性假设，主要依赖于依存关系实现三元组实例的抽取。

4）文献[24]　主要采用基于 Adaboost 迭代算法的协同训练方法对关系抽取模型进行强化，以缓解三元组实例含有噪声和错误的问题。其中，文本语句中的实体标记工作依赖于 Stanford CoreNLP 系统。

5）文献[25]　主要通过远程监督（distant supervision）实体关系抽取方法，利用 freebase 知识库和 wikipedia 文本库自动获取关系抽取训练数据，并训练模型实现实体关系抽取任务。其中，文本语句中的实体标记工作依赖于 Stanford CoreNLP 系统。

6）本文方法（未做共指消解处理）（可简记为 MDCGCV1）基于本文模型框架构建概念图，但是在处理单篇文档时，并未对其中的共指代词做消解处理。

7）本文方法（未做事实过滤处理）（可简记为 MDCGCV2）基于本文模型框架，将 OLLIE 的抽取结果直接输入等价概念及关系合并的任务，进而构建概念图。

由表 8 中的比较结果可以看出，本文提出的模型分别以 84.0%的主题概念覆盖率、94.7%的事实知识兼容性以及 0.474 的 ROUGE F1 值优胜于其他对比方法，说明了该模型在构建高质量概念图上的有效性。

表 8　代表性算法的实验结果比较

Table 8　Comparison of experimental results for representative algorithms

模型方法	主题概念覆盖率/%	等价概念对数量/n	新的语义关系标签数量/n	强连通分量个数/n	事实知识的兼容性/%	ROUGE-2/F1
本文方法	**84.0**	10	**3**	**3**	**94.7**	**0.474**
文献[21]	70.4	6	5	6	79.2	0.346
文献[22]	**80.6**	7	5	5	**92.1**	0.372
文献[23]	**81.6**	8	**3**	**3**	**91.6**	**0.470**
文献[24]	72.8	9	7	5	87.6	**0.442**
文献[25]	78.4	10	5	5	**91.7**	0.348
MDCGCV1	64.0	15	9	7	42.3	0.122
MDCGCV2	67.2	8	8	9	44.2	0.146

文献[21]模型主要利用句子的语言结构信息来抽取三元组实例。在主题概念覆盖率指标上，与本文模型相差 10 个左右的百分点，其主要原因在于是本文模型所使用的 OLLIE 算法能够更好地解决语句长程依赖问题，其精度相对较高；而文献[21]模型受到抽取精度的影响，其在事实知识的兼容性指标上的值为 79.2%，在 ROUGE-2 标准下的 F1 值为 0.346。

文献[22]模型的缺陷主要可归结于两个方面：a）单纯使用主谓宾句法关系进行三元组实例的抽取，对于长句的解析存在一些问题；b）过度依赖于正确的分词结果。文献[23]模型在文献[22]模型的基础上，对句法结构中的依存关系进行识别与分析。从实验结果来看，其整体性能优于前者，主题概念的覆盖率达到了 81.6%。文献[22]和[23]两个模型在事实知识的兼容性指标上均超过了 90%，由此也可以看出以句法分析为基础，本文所提出的事实过滤算法的有效性。

文献[24]模型主要采用基于 Adaboost 迭代算法的协同训练

方法对关系抽取模型进行强化，在一定程度上能够缓解三元组实例的噪声问题，但对于本实验语料中同一主题下的知识单元过于分散、语料完备性较差等问题，该模型的表现能力仍然有限，在概念覆盖率指标上的值为 72.8%。

文献[25]模型采用远程监督的方法实现关系抽取任务，然而其在主题概念的覆盖率指标上的值为 78.4%，分析其原因主要是因为：a）该方法在标注数据的获取过程中，主要借助于以下假设：所有包含实体对的句子都蕴涵了两者之间潜在的关系，而本实验所提供的文本语料并不完备，未能较好地支持上述假设；b）在本实验的语料中，对于未指定的关系类型[26]未能实现较好的标注；c）训练数据的样本过少，该模型的性能受到制约。

MDCGCV1 模型并未对单篇文档中的共指代词做消解处理，导致：a）指代不清或无效的三元组实例的数量显著增加；b）抽取到的事实知识的可读性较差；c）受到事实知识中语义单元模糊性的影响，通过事实过滤算法而产生的三元组实例出现“假兼容”的现象。因此，该模型的平均主题概念覆盖率只有 64.0%；事实知识的兼容性只有 42.3%，低于所有方法中该项指标的平均值。MDCGCV2 模型中，由于 OLLIE 模型的抽取结果未经过事实过滤流程，受该模型本身精度以及误差传播的影响，无信息量或冗余的三元组实例无法得到有效的处理，MDCGCV2 模型的主题概念覆盖率以及事实知识的兼容性分别为 67.2%与 44.2%。

3　结束语

为了解决主题信息跨文档分布，用户难以从中挖掘并组织核心概念以及语义连接的问题，本文提出了一个基于开放域抽取的多文档概念图构建模型。为了验证模型的有效性，在 Signal Media 发布的真实新闻数据集上就主题概念覆盖率、概念图连通性、概念图可读性、模型运行时间以及对比算法五个方面进行了实验验证，实验结果表明，本文提出的概念图构建模型能够跨文档挖掘并组织与特定主题相关的关键信息，并通过概念图表现其中显著的实体、概念，以及它们之间的关系。概念图在主题概念覆盖率、事实知识的兼容性等指标上均取得了较好的效果；除此之外，其对于自动文档摘要的应用也具有重要的参考价值。但是本文模型仍存在一定的局限性，如概念及关系的抽取任务主要依赖于开放域抽取系统 OLLIE、抽取出的三元组实例含有较大的噪声。除此之外，OLLIE 仅限于英文文本，无法应用于结构复杂且包含多语义概念的中文文本语料。因此，下一步将尝试将语义依存分析引入到本研究中，从文档的主题句中更加精准地提取主题词对以及它们之间的语义关系，并尝试进一步扩大本文模型的适用范围。

参考文献：

[1] Novak J D, Cañas A J. Theoretical origins of concept maps, how to construct them, and uses in education [J]. Reflecting Education, 2007, 3 (1): 29-42.

[2] Erkan, Günes, Dragomir R. Radev. LexRank: graph-based lexical centrality as salience in text summarization [J]. Journal of Artificial Intelligence Research, 2004, 22: 457-479.

[3] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks [C]// Advances in Neural Information Processing Systems. 2014: 3104-3112.

[4] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3 (1): 993-1022.

[5] Celikyilmaz A, Hakkani-Tur D. A hybrid hierarchical model for multi-document summarization [C]// Proc of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2010: 815-824.

[6] Lin C Y, Hovy E H. From single to multi-document summarization [C]// Proc of the 40th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2002: 457-464.

[7] Fabbrizio G D, Aker A, Gaizauskas R. Summarizing online reviews using aspect rating distributions and language modeling [J]. IEEE Intelligent Systems, 2013, 28 (3): 28-37.

[8] Banko M, Cafarella M J, Soderland S, *et al.* Open information extraction from the Web [C]// Proc of the 18th International Joint Conference on Artifical Intelligence. Cambridge, MA: AAAI, 2007: 2670-2676.

[9] 杨博，蔡东风，杨华. 开放式信息抽取研究进展 [J]. 中文信息学报, 2014, 28 (4): 1-11. (Yang Bo, Cai Dongfeng, Yang Hua. Progress in open information extraction [J]. Journal of Chinese Information Processing, 2014, 28 (4): 1-11.)

[10] Manning C, Surdeanu M, Bauer J, *et al.* Multi-document relationship fusion via constraints on probabilistic databases [C]// Proc of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. 2007: 332-339.

[11] 刘绍毓，李弼程，郭志刚，等. 实体关系抽取研究综述 [J]. 信息工程大学学报, 2016, 17 (5): 541-547. (Liu Shaoyu, Li Bicheng, Guo Zhigang, *et al.* Review of entity relation extraction [J]. Journal of Chinese Information Processing, 2016, 17 (5): 541-547.)

[12] 王林玉，王莉，郑婷一. 基于卷积神经网络和关键词策略的实体关系抽取方法 [J]. 模式识别与人工智能, 2017, 30 (5): 465-472. (Wang Linyu, Wang Li, Zheng Tingyi. Entity relation extraction based on convolutional neural network and keywords strategy [J]. Pattern Recognition and Artificial Intelligence, 2017, 30 (5): 465-472.)

[13] Zeng D, Liu K, Chen Y, *et al.* Distant supervision for relation extraction via piecewise convolutional neural networks [C]// Proc of Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015: 1753-1762.

[14] Kumar S. A survey of deep learning methods for relation extraction [EB/OL]. (2017) . https://arxiv. org/abs/1705. 03645.

[15] Manning C, Surdeanu M, Bauer J, *et al.* The stanford CoreNLP natural language processing Toolkit [C]// Proc of the 52nd Annual Meeting of the



Association for Computational Linguistics: System Demonstrations. Stroudsburg: Association for Computational Linguistics, 2014: 55-60.

[16] Mihalcea R, Tarau P. TextRank: bringing order into texts [C]// Proc of Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2004: 404-411.

[17] Pilehvar M T, Jurgens D, Navigli R. Align, disambiguate and walk: a unified approach for measuring semantic similarity [C]// Proc of the 51st Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2013: 1341-1351.

[18] 王桐，王磊，吴吉义，等. WordNet 中的综合概念语义相似度计算方法 [J]. 北京邮电大学学报, 2013, 36 (2): 98-101. (Wang Tong, Wang Lei, Wu Jiyi, *et al.* Semantic similarity calculation method of comprehensive concept in WordNet [J]. Journal of Beijing University of Posts and Telecommunications, 2013, 36 (2): 98-101.)

[19] 文本理解会议标准语料库 [EB/OL]. [2014-09-09]. https://duc. nist. gov/. (Text understanding conference standard corpus [EB/OL]. [2014-09-09]. https://duc. nist. gov/.)

[20] Li Jingxuan, Li Lei, Li Tao. Multi-document summarization via submodularity [J]. Applied Intelligence, 37 (3): 420-430.

[21] Angeli G, Premkumar M J J, Manning C D. Leveraging linguistic structure for open domain information extraction [C]// Proc of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015: 344-354.

[22] Jing Tao, Zuo Wanli, Sun Jigui, *et al.* Semantic annotation of Chinese Web pages: from sentences to RDF representations [J]. Journal of Computer Research and Development, 2008, 45 (7): 1221-1231. 靓

[23] 荆涛. 面向领域网页的语义标注若干问题研究 [D]. 长春：吉林大学, 2011. (Jing Tao. Research on semantic annotation of domain oriented Web pages [D]. Changchun: Jilin University, 2011.) .

[24] 王旭阳，姜喜秋. 特定领域概念属性关系抽取方法研究 [J]. 吉林大学学报：信息科学版, 2017, 35 (4): 430-437. (Wang Xuyang, Jiang Xiqiu. research on extraction method of specific domain concept and property [J]. Journal of Jilin University: Information Science Edition, 2017, 35 (4): 430-437.)

[25] Mintz M, Bills S, Snow R, *et al.* Distant supervision for relation extraction without labeled data [C]// Proc of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Stroudsburg: Association for Computational Linguistics, 2009: 1003-1011.

[26] Chan Y S, Roth D. Exploiting background knowledge for relation extraction [C]// Proc of the 23rd International Conference on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2010: 152-160.